Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision

Year: 2023 | *Citations: 391* | *Authors: Zhiqing Sun, Yikang Shen, Qinhong Zhou*

Abstract

Recent AI-assistant agents, such as ChatGPT, predominantly rely on supervised fine-tuning (SFT) with human annotations and reinforcement learning from human feedback (RLHF) to align the output of large language models (LLMs) with human intentions, ensuring they are helpful, ethical, and reliable. However, this dependence can significantly constrain the true potential of AI-assistant agents due to the high cost of obtaining human supervision and the related issues on quality, reliability, diversity, self-consistency, and undesirable biases. To address these challenges, we propose a novel approach called SELF-ALIGN, which combines principle-driven reasoning and the generative power of LLMs for the self-alignment of AI agents with minimal human supervision. Our approach encompasses four stages: first, we use an LLM to generate synthetic prompts, and a topic-guided method to augment the prompt diversity; second, we use a small set of human-written principles for AI models to follow, and guide the LLM through in-context learning from demonstrations (of principles application) to produce helpful, ethical, and reliable responses to user's queries; third, we fine-tune the original LLM with the high-quality self-aligned responses so that the resulting model can generate desirable responses for each query directly without the principle set and the demonstrations anymore; and finally, we offer a refinement step to address the issues of overly-brief or indirect responses. Applying SELF-ALIGN to the LLaMA-65b base language model, we develop an AI assistant named Dromedary. With fewer than 300 lines of human annotations (including<200 seed prompts, 16 generic principles, and 5 exemplars for in-context learning). Dromedary significantly surpasses the performance of several state-of-the-art AI systems, including Text-Davinci-003 and Alpaca, on benchmark datasets with various settings.